

SYSTEM AND METHOD FOR IDENTIFYING COMPOUNDS THROUGH ITERATIVE ANALYSIS

Abstract

A system and method for identifying compounds through iterative analysis
5 of measure of association is disclosed. A limit on a number of tokens per
compound is specified. Compounds within a text corpus are iteratively evaluated.
A number of occurrences of one or more n -grams within the text corpus is
determined. Each n -gram includes up to a maximum number of tokens, which are
each provided in a vocabulary for the text corpus. At least one n -gram including a
10 number of tokens equal to the limit based on the number of occurrences is
identified. A measure of association between the tokens in the identified n -gram
is determined. Each identified n -gram with a sufficient measure of association is
added to the vocabulary as a compound token and the limit is adjusted.